Supplementary Materials: PreLAR: World Model Pre-training with Learnable Action Representation

Lixuan Zhang[1,2], Meina Kan[1,2], Shiguang Shan[1,2,3], and Xilin Chen[1,2]

[1] Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Peng Cheng Laboratory, Shenzhen, 518055, China
lixuan.zhang@vipl.ict.ac.cn, {kanmeina, sgshan, xlchen}@ict.ac.cn

This supplementary document provides additional information to clarify the principle and implementation of our scheme PreLAR, including the behavior learning details (Sec. 1), derivations of the *log-likelihood maximization loss* (Sec. 2), the algorithm of the world model pre-training and fine-tuning in PreLAR (Sec. 3), the qualitative results to show the learned action representation (Sec. 4), and the implementation details of the PreLAR (Sec. 5).

1 Behavior Learning

To learn the policy to complete a specific visual control task based on the fine-tuned world model from Eq. (12), we use the actor-critic reinforcement learning algorithm for behavior learning. This approach involves the use of the actor-critic scheme, which operates on imaginary latent trajectories $\{\hat{s}_t, \hat{a}_t, \hat{r}_t\}_{t=1}^{H}$ over a horizon H, similar to the methodology adopted in DreamerV2 [3]:

$$\begin{aligned} &\text{Actor:} \quad \hat{a}_t \sim p_\psi(\hat{a}_t|\hat{s}_t), \\ &\text{Critic:} \quad v_\xi(\hat{s}_t) \approx \mathbb{E}_{p_\theta, p_\phi}\left[\textstyle\sum_{i\le t} \gamma^{i-t}\hat{r}_i\right]. \end{aligned} \tag{1}$$

The actor and critic models are detailed in Eq. (13), wherein the critic is trained through regression on the λ-target cumulative rewards, as described in [5]:

$$\mathcal{L}(\xi) \doteq \mathbb{E}_{p_\theta, p_\psi}\left[\sum_{t=1}^{H-1} \tfrac{1}{2}\left(v_\xi(\hat{s}_t) - \mathtt{sg}(V_t^\lambda)\right)^2\right], \tag{2}$$

where the $\mathtt{sg}$ is the stop gradient function, and the λ-return is defined as:

$$V_t^\lambda \doteq \hat{r}_t + \gamma \begin{cases} (1-\lambda) v_\xi(\hat{s}_{t+1}) + \lambda V_{t+1}^\lambda & \text{if } t < H \\ v_\xi(\hat{s}_H) & \text{if } t = H. \end{cases} \tag{3}$$

The actor is trained to maximize the imagined cumulative rewards via the straight-through estimator [1] for propagating the value gradient through the world model with an entropy regularization loss to encourage exploration:

$$\mathcal{L}(\psi) \doteq \mathbb{E}_{p_\theta, p_\psi}\left[-V_t^\lambda - \eta\, \mathrm{H}\left[a_t|\hat{s}_t\right]\right]. \tag{4}$$

2 Derivations of the log-likelihood maximization loss

In this section, we derive the *log-likelihood maximization loss* in Eq. (5) of the body text. The derivation of the variational bound for *log-likelihood maximization loss* follows Jensen's Inequality as follows:

$$\begin{aligned}
\ln p(o_{1:T}) &\triangleq \ln \int p(o_{1:T}|\tilde{a}_{1:T})p(\tilde{a}_{1:T})d\tilde{a}_{1:T} \\
&= \ln \int \frac{p(o_{1:T}|\tilde{a}_{1:T})p(\tilde{a}_{1:T})}{q(\tilde{a}_{1:T}|o_{1:T})} q(\tilde{a}_{1:T}|o_{1:T})d\tilde{a}_{1:T} \\
&= \ln \mathbb{E}_{q(\tilde{a}_{1:T}|o_{1:T})}[\frac{p(o_{1:T}|\tilde{a}_{1:T})p(\tilde{a}_{1:T})}{q(\tilde{a}_{1:T}|o_{1:T})}] \\
&\geq \mathbb{E}_{q(\tilde{a}_{1:T}|o_{1:T})}[\ln \frac{p(o_{1:T}|\tilde{a}_{1:T})p(\tilde{a}_{1:T})}{q(\tilde{a}_{1:T}|o_{1:T})}] \\
&= \mathbb{E}_{q(\tilde{a}_{1:T}|o_{1:T})}[\ln p(o_{1:T}|\tilde{a}_{1:T})] - \mathrm{KL}[q(\tilde{a}_{1:T}|o_{1:T})\|p(\tilde{a}_{1:T})],
\end{aligned} \tag{5}$$

where the derivations of the *log-likelihood of action conditional video data* (the first term in Eq. (5)) refers to [2] as follows:

$$\begin{aligned}
\mathbb{E}_{q_\phi(\tilde{a}_{1:T}|o_{1:T})}[\ln p(o_{1:T}|\tilde{a}_{1:T})] = \mathbb{E}_{q_\phi(s_{1:T}|o_{1:T})q_\phi(\tilde{a}_{1:T}|o_{1:T})}\Big[\sum\nolimits_{t=1}^{T}\Big(\underbrace{\ln p_\phi(o_t|s_t)}_{\text{image log loss}} \\
\underbrace{-\beta \mathrm{KL}[q_\phi(s_t|s_{t-1},\tilde{a}_{t-1},o_t)\|p_\phi(\hat{s}_t|s_{t-1},\tilde{a}_{t-1})]}_{\text{dynamics KL loss}}\Big)\Big].
\end{aligned} \tag{6}$$

Based on the assumption that the current action solely depends on the observations in the current and subsequent time steps, the *action Kullback–Leibler (KL)* (the second term in Eq. (5)) is decomposed according to the temporal as follows:

$$\begin{aligned}
&\mathrm{KL}[q(\tilde{a}_{1:T}|o_{1:T})\|p(\tilde{a}_{1:T})] \\
&= \int q(\tilde{a}_{1:T}|o_{1:T}) \ln \frac{q(\tilde{a}_{1:T}|o_{1:T})}{p(\tilde{a}_{1:T})} d\tilde{a}_{1:T} \\
&= \int q(\tilde{a}_1|\tilde{a}_{2:T},o_{1:T})q(\tilde{a}_2|\tilde{a}_{3:T},o_{1:T})\cdots q(\tilde{a}_T|o_{1:T})\cdot \\
&\quad \ln \frac{q(\tilde{a}_1|\tilde{a}_{2:T},o_{1:T})q(\tilde{a}_2|\tilde{a}_{3:T},o_{1:T})\cdots q(\tilde{a}_T|o_{1:T})}{p(\tilde{a}_1)p(\tilde{a}_2)\cdots p(\tilde{a}_T)} d\tilde{a}_1 d\tilde{a}_2 \cdots d\tilde{a}_T \\
&= \int q(\tilde{a}_1|o_1,o_2)q(\tilde{a}_2|o_2,o_3)\cdots q(\tilde{a}_T|o_T)\cdot \\
&\quad \ln \frac{q(\tilde{a}_1|o_1,o_2)q(\tilde{a}_2|o_2,o_3)\cdots q(\tilde{a}_T|o_T)}{p(\tilde{a}_1)p(\tilde{a}_2)\cdots p(\tilde{a}_T)} d\tilde{a}_1 d\tilde{a}_2 \cdots d\tilde{a}_T \\
&= \int q(\tilde{a}_1|o_1,o_2)q(\tilde{a}_2|o_2,o_3)\cdots q(\tilde{a}_T|o_T)\cdot \\
&\quad \left(\ln \frac{q(\tilde{a}_1|o_1,o_2)}{p(\tilde{a}_1)} + \ln \frac{q(\tilde{a}_2|o_2,o_3)}{p(\tilde{a}_2)} + \cdots + \ln \frac{q(\tilde{a}_T|o_T)}{p(\tilde{a}_T)}\right) d\tilde{a}_1 d\tilde{a}_2 \cdots d\tilde{a}_T.
\end{aligned} \tag{7}$$

The integration of the t-th term in round brackets on the last line of Eq. (7) is derived as follows:

$$\begin{aligned}
&\int q(\tilde{a}_1|o_1,o_2)q(\tilde{a}_2|o_2,o_3)\cdots q(\tilde{a}_T|o_T)\ln\frac{q(\tilde{a}_t|o_t,o_{t+1})}{p(\tilde{a}_t)}d\tilde{a}_1 d\tilde{a}_2\cdots d\tilde{a}_T \\
=&\int q(\tilde{a}_1|o_1,o_2)\cdots q(\tilde{a}_{t-1}|o_{t-1},o_t)q(\tilde{a}_{t+1}|o_{t+1},o_{t+2})\cdots q(\tilde{a}_T|o_T)\cdot \\
&\left[\int q(\tilde{a}_t|o_t,o_{t+1})\ln\frac{q(\tilde{a}_t|o_t,o_{t+1})}{p(\tilde{a}_t)}d\tilde{a}_t\right]d\tilde{a}_1\cdots d\tilde{a}_{t-1}d\tilde{a}_{t+1}\cdots d\tilde{a}_T \\
=&\,\mathrm{KL}[q(\tilde{a}_t|o_t,o_{t+1})\|p(\tilde{a}_t)]\cdot \\
&\int q(\tilde{a}_1|o_1,o_2)\cdots q(\tilde{a}_{t-1}|o_{t-1},o_t)q(\tilde{a}_{t+1}|o_{t+1},o_{t+2})\cdots q(\tilde{a}_T|o_T)\cdot \\
&d\tilde{a}_1\cdots d\tilde{a}_{t-1}d\tilde{a}_{t+1}\cdots d\tilde{a}_T \\
=&\,\mathrm{KL}[q(\tilde{a}_t|o_t,o_{t+1})\|p(\tilde{a}_t)].
\end{aligned} \tag{8}$$

Applying Eq. (8) to Eq. (7), the final action KL loss term is as follows:

$$\mathrm{KL}[q(\tilde{a}_{1:T}|o_{1:T})\|p(\tilde{a}_{1:T})]=\sum_{t=1}^{T}\mathrm{KL}[q(\tilde{a}_t|o_t,o_{t+1})\|p(\tilde{a}_t)]. \tag{9}$$

Here we let $q(\tilde{a}_T|o_T,o_{T+1})\triangleq q(\tilde{a}_T|o_T)$ for concision. In the implementation, the T-th term is ignored for correctness. In summary, the final *log-likelihood maximization loss* (i.e. Eq. (5) in the body text) is to minimize the objective as follows:

$$\begin{aligned}
\mathcal{L}^{\texttt{like}}(\phi)=\mathbb{E}_{q_\phi(s_{1:T}|o_{1:T})q_\phi(\tilde{a}_{1:T}|o_{1:T})}\Big[\sum\nolimits_{t=1}^{T}\Big(&\underbrace{-\ln p_\phi(o_t|s_t)}_{\text{image log loss}} \\
&\underbrace{+\beta\mathrm{KL}[q_\phi(s_t|s_{t-1},\tilde{a}_{t-1},o_t)\|p_\phi(\hat{s}_t|s_{t-1},\tilde{a}_{t-1})]}_{\text{dynamics KL loss}}\Big)\Big]\ \underbrace{+\beta_a\sum_{t=1}^{T}\mathrm{KL}[q_\phi(\tilde{a}_t|o_t,o_{t+1})\|p(\tilde{a}_t)]}_{\text{action KL loss}}.
\end{aligned} \tag{10}$$

3 Algorithm of the world model pre-training and fine-tuning

Following the similar pre-training and fine-tuning procedures as APV [4] and ContextWM [6], the pre-training and fine-tuning procedures of PreLAR are illustrated in Algorithm 1. The world model in the pre-training and fine-tuning phases are summarised in Eq. (11) and Eq. (12) as follows:

$$\begin{aligned}
&\text{Representation model:} && s_t\sim q_\phi(s_t|s_{t-1},\tilde{a}_{t-1},o_t), \\
&\text{Transition model:} && \hat{s}_t\sim p_\phi(\hat{s}_t|s_{t-1},\tilde{a}_{t-1}), \\
&\text{Image decoder:} && \hat{o}_t\sim p_\phi(\hat{o}_t|s_t), \\
&\text{Inverse dynamics encoder:} && \tilde{a}_t\sim q_\phi(\tilde{a}_t|o_t,o_{t+1}).
\end{aligned} \tag{11}$$

Algorithm 1 PreLAR: World Model Pre-training with Learnable Action Representation

// **World model pre-training with learnable representation**
Random initialize the parameters ϕ of the action-conditional dynamics model (representation model, transition model, image decoder, and inverse dynamics encoder)

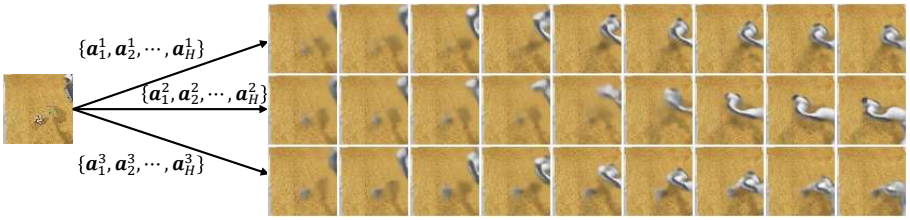
Load action-free video dataset $\mathcal{D}$
for each training step **do**
  // World model pre-training
  Random sample minibatch $\{(o_t)_{t=1}^{T}\} \sim \mathcal{D}$
  Update world model by minimizing $\mathcal{L}^{\texttt{pre-train}}(\phi)$ in Eq. (14)
end for

// **World fine-tuning with real action**
Load pre-trained parameters ϕ of action-conditional dynamics model (representation model, transition model, image decoder)
Random initialize parameters θ of action encoder and reward predictor
Random initialize parameters ψ, ξ of actor and critic
Random initialize replay buffer $\mathcal{B}$ with random exploration
for each timestep t **do**
  // Collect trajectories
  Get representation $\tilde{a}_{t-1} \sim p_\theta(\tilde{a}_t|a_t), s_t \sim q_\phi(s_t|s_{t-1}, \tilde{a}_{t-1}, o_t)$
  Get action $a_t \sim p_\psi(a_t|s_t)$
  Add trajectory $\{o_t, a_t, r_t\}$ to replay buffer $\mathcal{B}$
  // World model fine-tuning
  Random sample minibatch $\{o_t, a_t, r_t\}_{t=1}^{T} \sim \mathcal{B}$
  Update world model by minimizing $\mathcal{L}^{\texttt{fine-tune}}(\phi, \theta)$ in Eq. (15)
  // Behavior learning
  Imagine trajectory $\{\hat{s}_t, \hat{a}_t, \hat{r}_t\}_{t=1}^{H}$ using world model and actor
  Update actor and critic by minimizing objectives $\mathcal{L}(\psi)$ in Eq. (16) and $\mathcal{L}(\xi)$ in Eq. (17)
end for

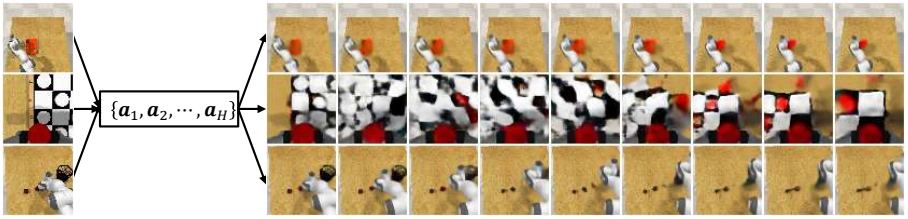
$$
\begin{aligned}
&\text{Representation model:} && s_t \sim q_\phi(s_t|s_{t-1}, \tilde{a}_{t-1}, o_t),\\
&\text{Transition model:} && \hat{s}_t \sim p_\phi(\hat{s}_t|s_{t-1}, \tilde{a}_{t-1}),\\
&\text{Action encoder:} && \tilde{a}_t \sim p_\theta(\tilde{a}_t|a_t),\\
&\text{Image decoder:} && \hat{o}_t \sim p_\phi(\hat{o}_t|s_t),\\
&\text{Reward predictor:} && \hat{r}_t \sim p_\theta(\hat{r}_t|s_t).
\end{aligned}
\tag{12}
$$

The actor and critic models are detailed in Eq. (13) as follows:

$$
\begin{aligned}
&\text{Actor:} && \hat{a}_t \sim p_\psi(\hat{a}_t|\hat{s}_t),\\
&\text{Critic:} && v_\xi(\hat{s}_t) \approx \mathbb{E}_{p_\theta, p_\phi}\left[\textstyle\sum_{i \le t} \gamma^{i-t} \hat{r}_i\right].
\end{aligned}
\tag{13}
$$



(a) Same initial observation, condition on three different actions.



(b) Three different initial observations, condition on the same action.

Fig. 1: Qualitative results of the learned world model.

The objectives to train the pre-training models (Eq. (14)), fine-tuning models (Eq. (15)), actor (Eq. (16)), and critic (Eq. (17)) models are brought here for convenience, as follows:

$$\mathcal{L}^{\texttt{pre-train}}(\phi) = \mathcal{L}^{\texttt{like}}(\phi) + \beta_r \mathcal{L}^r(\phi), \tag{14}$$

$$\mathcal{L}^{\texttt{fine-tune}}(\phi, \theta) = \mathbb{E}_{q_\phi(s_{1:T}|o_{1:T})p_\theta(\tilde{a}_{1:T}|a_{1:T})}\Big[\sum_{t=1}^{T}\Big(\underbrace{-\ln p_\phi(o_t|s_t)}_{\text{image log loss}}\underbrace{-\ln p_\theta(r_t|s_t)}_{\text{reward log loss}}$$
$$\underbrace{-\ln p_\theta(r_t + r_t^{\texttt{int}}|s_t)}_{\text{auxiliary reward log loss}}\underbrace{+\beta \mathrm{KL}[q_\phi(s_t|s_{t-1}, \tilde{a}_{t-1}, o_t)\|p_\phi(\hat{s}_t|s_{t-1}, \tilde{a}_{t-1})]}_{\text{dynamics KL loss}}\Big)\Big], \tag{15}$$

$$\mathcal{L}(\psi) \doteq \mathbb{E}_{p_\theta, p_\psi}\left[-V_t^\lambda - \eta\, \mathrm{H}\left[a_t|\hat{s}_t\right]\right], \tag{16}$$

$$\mathcal{L}(\xi) \doteq \mathbb{E}_{p_\theta, p_\psi}\left[\sum_{t=1}^{H-1}\frac{1}{2}\left(v_\xi(\hat{s}_t) - \mathtt{sg}(V_t^\lambda)\right)^2\right]. \tag{17}$$

4 Qualitative Results

We present the qualitative results from two aspects as shown in Fig. 1: (a) *Given the same initial observation, sample three action representation sequences.* The generated observations differ, appearing as slight left movement, more pronounced left movement, and downward movement, respectively. (b) *Given three different initial observations, sample the same action representation sequence.*

The generated observation sequences exhibit similar trends, all appearing to move toward the upper right. These qualitative results clearly indicate that the observations generated by the model are coherent with the conditional action information, and the latent actions (i.e. action representation) encode the manipulator motion information qualitatively.

5 Implementation Details

We utilize the same hyperparameters and implementation with APV [4] and ContextWM [6]. The newly added hyperparameters and architectures have been explained in Sec. 4.1 of the body text. Here we report the important hyperparameters in Tab. 1 for completeness.

Moreover, we use 8 × GeForce RTX 3090 in our experiments. The model parameters are approximately 27.35 M in the pre-training phase and 27.77 M in the fine-tuning phase. The inference frequency of our world model in GeForce RTX 3090 is about 167 Hz.

References

1. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
2. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels. In: International Conference on Machine Learning (ICML). pp. 2555–2565 (2019)
3. Hafner, D., Lillicrap, T.P., Norouzi, M., Ba, J.: Mastering atari with discrete world models. In: International Conference on Learning Representations (ICLR) (2021)
4. Seo, Y., Lee, K., James, S.L., Abbeel, P.: Reinforcement learning with action-free pre-training from videos. In: International Conference on Machine Learning (ICML). pp. 19561–19579 (2022)
5. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. Robotica **17**(2), 229–235 (1999)
6. Wu, J., Ma, H., Deng, C., Long, M.: Pre-training contextualized world models with in-the-wild videos for reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 39719–39743 (2023)

Table 1: The hyperparameters in experiments. The most hyperparameters are the same as APV [4] and ContextWM [6] if without additional explanation.

	Hyperparameter	Value
pre-training phase	Image size	$64 \times 64 \times 3$
	Image preprocess	Linearly rescale from $[0, 255]$ to $[-0.5, 0.5]$
	Video segment length T	25
	dynamics KL weight β	1
	Optimizer	Adam
	Learning rate	3×10^{-4}
	Batch size	16
	Training iterations	6×10^5
	action-sate consistency loss weight β_r (new)	1
	action KL weight β_a (new)	1
fine-tuning phase	Observation size	$64 \times 64 \times 3$
	Image preprocess	Linearly rescale from $[0, 255]$ to $[-0.5, 0.5]$
	Trajectory segment length T	50
	Random exploration	5000
	Replay buffer capacity	1×10^6
	Training frequency	Every 5 environment steps
	dynamics KL weight β	1
	Imagination horizon H	15
	Discount γ	0.99
	λ-target discount	0.95
	Entropy regularization η	1×10^{-4}
	Batch size	50
	World model optimizer	Adam
	World model learning rate	3×10^{-4}
	Actor optimizer	Adam
	Actor learning rate	8×10^{-5}
	Critic optimizer	Adam
	Critic learning rate	8×10^{-5}
	Evaluation episode	10